人工知能と対話技法 26－1
（1982. 6. 11）

発見的手法による知識ベースの検索

志村正道

（東京工業大学 工学部 情報工学科）

§1. まえがき

近年，知識を利用するエキスパートシステムなどが種々開発され，その実用化も盛んになりつつある。本論文で報告するシステムは植物の名前を同定するための検索システムである。一般に，植物図鑑においては植物名からその植物の特徴を調べるのは容易であるが，その逆は極めて困難である。ここでは，植物の特徴からその植物名を推定する検索システムについて考える。

検索システムはいくつかのキーワードによって，その検索を行なうのが基本的である。すなわち，木構造状の検索項目を木の探索によって検索する方法が基本的である。しかし，実際には意味のない冗長なキーワードがあるとか，検索項目が多い場合などには，このような木の探索方法は効率がよくない。そのため，探索において発見的情報を用いた手法が種々報告されている。ところが，そのようなシステムでは知識ベースとしてテーブルを用いた場合，実際にはキーワードが欠除していたり，あるいはキーワードを変更したりすることが必ずしも容易ではないという欠点が生ずる。したがって，知識をプロダクションルールの形で表現することを考え，実際に植物を対象とする検索システムを試作した。本論文ではこのシステムについて述べ，またその実験結果について報告する。

§2. 知識の表現

知識の追加・修正・消去・確認が容易に行なえるような形式を考える。樹木状に知識を格納する方法が一般的であるが，次のような欠点をもっている。すなわち，

(i) キーワードが欠損している場合探索をすすめることが困難である。
(ii) データ構造が効率的でなくなる。
(iii) 知識あるいはデータの拡張が困難である。
(iv) 多くのデータを扱い難い。

以上の理由によってここではプロダクションルール（PR）を用いたプロダクションシステムを用いることを考える。

いま，植物名をP_i，質問事項をQ_j，その回答をA_{jk}とするとき，Q_jに関するP_iの知識を

$$P_i = Q_j\ (A_{jk})$$

と表わす。また複数の知識を表現するには

$$Q_j(A_{jk_1}) \cup Q_j(A_{jk_2}) \cup \cdots \cup Q_j(A_{jk_m})$$

ということではなく

$$Q_i\,(A_{ik_1} \cup A_{ik_2} \cup \cdots \cup A_{ik_m})$$

という知識表現を用いる。

また，葉の長さなどの数値をとり扱う場合にはmin値とmax値のみを知識としてもつことにし，Q'_i (min－max) という形で記憶する。

したがって，本システムで取り扱う知識は次のような二つの形である。

$$P = Q\ (A)$$
$$P = Q'\ (\min - \max)$$

このような知識は実際に内部コードに変換されて格納されるが，本システムにおいては図1に示すような表現形式を採用している。このような表現形式の特長として次のような点がある。例えば，まきひげの長さに関する質問事項はまきひげをもつ植物の検索には有効であるが，まきひげをもつ植物は全体からみればごく一部であって，それ以外の多くの植物に対しては無意味な質問となってしまう。したがって，関係ある植物のみに限ってルールを設定しておけばよく，2次元配列を用いる形式に比べて効率的な記憶領域の使用ができる。

P_1	Q_1	A_{11}	*
P_1	Q_2	A_{12}	*
P_1	Q_3	Min	Max
P_2	Q_2	A_{21}	*
⋮	⋮	⋮	⋮
P_i	Q_j	A_{ik}	*
⋮	⋮	⋮	⋮

図　1

§3.　知識の削除，追加，修正，合成

知識ベースを考える場合に重要なことの一つは知識の削除，追加，修正，合成であり，これらの機能が容易に実行できるようなシステムが望ましい。このようなシステムにおいては，とくに知識の量を増やして，知識ベースを拡張していくということは，本質的に要求される機能なのである。

上に述べたような機能を考えるには，実際に次のような問題を解決しておかねばならない。すなわち，

(i) $P=Q(A_1)$ なる知識がすでにあるか否を見出さなければならない。

(ii) $P=Q(A_1)$ に対し $P=Q(A_2)$ なるデータが入ったとき，新しい知識として $P=Q(A_1)$ を捨てて $P=Q(A_2)$ のみにするか，あるいは $P=Q(A_1 \cup A_2)$ のように両方とも正しいものとすべきかを決定しなければならない。本システムでは問合せが出来るようになっている。

(iii) 新しい植物名が入ったとき，過去の植物名の領域に新しい分割が入ってくる可能性があり，誤って検索されるとかあるいは検索効率が下がる副作用が生じないようにしなければならない。

§4. 知識データの検索

本システムの検索動作について述べる。対象の検索としては，検索される対象のもつ特徴をすべてシステムに入れ，これらの情報によって一度に行なう方法が考えられる。しかし，検索される対象によってそれぞれ特徴が異なるため効率が必ずしも良くはない。それゆえ，ここで述べるシステムにおいては，検索するのにまず最も適当と考えられる特徴を求め，これについてシステムが検索者に質問を発する。検索者はこの質問に回答するという形をとりながら，逐次特徴に関する情報をシステムに取り入れる。システムは入力された情報すなわち特徴にしたがって検索を行なうが，この動作は探索対象が一意に定まるまで続けられる。

さて，プロダクションシステムにおける認識－作動サイクルは次の3つのステップから構成されている。

(1) 知識データ（プロダクション）の中から起動条件の満たされているものの集合をつくる。
(2) その集合の中から，何らかの方法で起動すべきプロダクションを選択する。
(3) 選択されたプロダクションに従って，実行を行なう。

本システムにおいては，プロダクションルールの起動は質問の発生ということに相当するが，このとき，質問事項の選択は次のような定め方が考えられる。

(1) 固定した順序で質問をする。
(2) 最適な順序で質問をする。
(3) 最も新しく使われた順序で質問をする。

このうち(2)の最適な順序とは，その質問事項がなるべく多くの植物に関連しており，しかもなるべく多くの分類が可能な質問から選んでいった順序のことである。また，質問回答の結果より，順次条件に適合するプロダクションルールを起動させていき，それに合わない植物を棄却していく。

このような手順をスーパステップとよぶが，このスーパステップを実行した場合，検索は次のような状態に至ったときに終了する。

(1) 条件にあう植物が一つになった状態
(2) 質問をすべて出しつくした状態
(3) 候補植物がなくなった状態

すなわち，(1)の場合は検索を行なっていた対象が決定したわけであるから成功したことになる。(2)の場合は検索すべき入力情報が不足あるいは不備であった場合である。(3)は誤情報あるいは知識の不足のために検索対象が見つからなかった場合である。

§5. 知識ベースのデータ構造

検索システムにおいて用いられる種々のデータのデータ構造を以下に示す。

(i) ＪＯＨＯ

ＪＯＨＯは内部コードに変換された知識データを格納し，6次元（6 Wds）の配列である。各ワードの内容は 1. 植物名Ｐ， 2. 質問事項Ｑ， 3. 回答Ａ， 4. 最小値，最大値， 5. Ｐのポインタ， 6. Ｑのポインタである。これらの関係を示

すと図2のようになる。

1	P_1	Q_1	A_{11}	*	P_1のポインタ	Q_1のポインタ
2	P_2	Q_2	min	max	P_2のポインタ	Q_2のポインタ
	⋮	⋮	⋮	⋮	⋮	⋮
N	P_i	Q_j	A_{jk}	*	N''	N'
	⋮	⋮	⋮	⋮	⋮	⋮
N′	$P_{i'}$	$Q_{j'}$		*		
N*	$P_{i'}$	Q_j	$A_{jk'}$	*	N'	0

図2

(ii) MOZIRT

入力された文字列は内部コードに変換されるが，MOZIRTは内部コードと文字列を対応させるために，これらの関係を格納しておく配列である。このとき，図3に示すように，1ワードに4文字があてられ，各文字列は先頭にワード長を記入した後に格納される。

	3	'KITS'	'UNEA'	'ZAMI'	1	'PI '	

図3

(iii) NLIB

植物名に関する情報を格納しておく4 Wdsの配列である。NLIBの4 Wdsの内容は 1. MOZIRTに格納されている植物名の文字列を示すポインタ， 2. 質問事項を示すポインタ， 3，4. 配列JOHOで最初および最後に現れる番地のポインタである。

(iv) FLIB

質問事項に関する情報を格納しておく5 Wdsの配列である。FLIBの5 Wdsの内容は1～4. NLIBと同じ， 5. 質問事項の種類（数値あるいは分岐）である。

(V) XLIB

回答に関する情報を格納しておく3 Wds配列である。XLIBの3 Wdsの内容は 1. MOZIRTのポインタ 2. 回答項目のポインタ 3. 質問事項のポインタである。

以上，5つの配列によって知識ベースが構成されている。すなわち，JOHOはデータそのものを，NLIB，FLIB，XLIBは植物名P，質問事項Q，回答項目Aの索引を，またMOZIRTはこれらに登録された文字列を表わす辞書である。

§6. 検索植物名の決定

検索システムでは質問に対する回答結果から候補植物名を決定していくが，この過程について述べる。まず，最初においてはすべての植物が候補となっているが，各候補には荷重が割り当てられ，この荷重の最大なものが最終候補となる。すなわち，検索される植物の特徴が質問事項によって候補植物と一致する場合には荷重を増加し，一致しない場合には大巾に減少させる。また不明の場合には少し減少させる。このような荷重操作により，逐次候補植物の個数は減少していき，前に述べたような終了の状態になったとき終了する。

システムが入力情報をとり入れるために発生させる質問はなるべく最小の質問回数によって該当植物名が決定されるように選ばれなければならない。このためには，その特徴によって分類される対象が多くなるようなものを選ばなければならない。例えば，質問の回答結果に応じて，候補対象が2つのクラス，すなわち，その質問に対応する特徴をもつものともたないものとに分類されるものとしよう。対象の数をNとし，これを一つの特徴すなわち一回の質問により (a) 1と N−1個 (b) N/2 と N/2 個に分類される場合を考えてみる。

(a) の場合は平均

$$\frac{N+1}{2} - \frac{1}{N} \approx \frac{N}{2}$$

(b) の場合は平均

$$\log_2 N$$

回の質問により分類されることとなる。したがって，(b)のようになるべく多数個づつの分類がなされるような特徴について質問することが望ましい。あるいは最も多くの候補を否定するような質問であることが望ましいとも考えられるので，その質問の有効性を，否定される候補数の期待値で表わすことにしよう。なお，質問の回答としてm個のうちのただ一つが入力されるものとし，二つ以上の回答はないものとする。

まず，選択式質問についての期待値は次のように求められる。質問Qに対して回答A_jが得られる確率をP_j，該当回答がない確率をP_0，またこのような質問に対して候補対象Nのうち否定される対象の数をそれぞれ$\bar{n}_j$，$\bar{n}_0$とする。このとき，期待値は

$$E = \sum_{j=1}^{m} P_j \bar{n}_j + P_0 \bar{n}_0$$

ただし，mは選択分岐数である。ここで，質問Qに関するデータをもっている対象の個数をn，回答A_jをもつ対象の個数をn_jとすると $\bar{n}_j = n - n_j$ となる。

したがって

$$P_j = \frac{n_j}{N}$$

$$P_0 = \frac{N-n}{N}$$

が得られ，期待値は

$$E = \sum_{j=1}^{m} \frac{n_j}{N}(n-n_j) + \frac{N-n}{N} n$$

となる。

次に数値式回答の期待値を求める。この場合回答は数値となるから，その値を v_j とし，また，$min \leqq v_j \leqq max$ を満足する対象の数を n_i とする。ただし，v_j は離散値で，min から max までの値を等確率でとるものとする。このとき v_j をとる確率 P_j は

$$P_j = \frac{1}{N} \sum_k \frac{1}{max_k - min_k + 1}$$

$$(min_k \leqq v_j \leqq max_k)$$

となり，期待値は

$$E = \sum_{v_j} \frac{n-n_j}{N} \left\{ \sum_k \frac{1}{max_k - min_k + 1} \right\} + \frac{N-n}{N} n$$

となる。

以下，この期待値の例を述べる。

例１．（N－m）種の対象がその質問に無関係で，関係する対象はそれぞれ異なる回答をもっている場合

$$E = \sum_{j=1}^{m} \frac{m-1}{N} + \frac{N-m}{N} \cdot m = \frac{m}{N}(N-1)$$

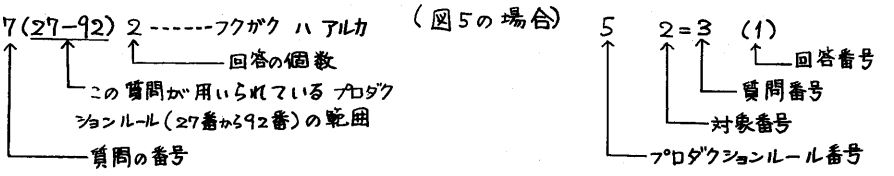
例２．　l 種の対象がその質問に無関係で，残り N－l 種が均等に m 個の回答をもっている場合

$$E = \frac{(N-l)^2}{N}\left(1 - \frac{1}{m}\right) + \frac{l}{N}(N-l)$$

実際のシステムにおける実験結果では，平均 7.5 個の質問事項によって候補植物名が決定されたのに対し，固定した一定順序の質問方法によれば 20～40 の質問を行なわねばならなかった。この結果から，上に述べた質問事項の選択はかなり有効であるといってよかろう。

§7. システムの概要

本システムの知識ベースに格納されている植物は 58 種である。また本システムで準備されている質問項目は図４に示すような 28 種である。図５はプロダクションルールの一例であって 467 個準備されている。図４および５における数字は下のような意味をもっている。



実際に検索を行なった例を図６に示しておく。

本システムにおける付随的ではあるが重要な機能に次のようなものがある。

(i) 検索中任意の時点で，候補植物名とその荷重を表示することができる。

(ii) 候補植物の決定に際し，使用したプロダクションルールを表示することができる。

(iii) 回答が不明の場合にも不明として入力できる。

```
1(  1-457) 2     ------ マキヒゲ ハ アルカ
                1       アル
                2       ナイ
2(  2-458) 2     ------ タクヨウ ハ アルカ
                1       アル
                2       ナイ
3(  3-462) 4     ------ タンヨウ カ フクヨウ カ
                1       タンヨウ
                2       5シュツ フクヨウ
                3       3シュツ フクヨウ
                4       ウジョウ フクヨウ
4( 12-463) 4     ------ ハ ノ ツキカタ ハ
                1       コンセイ
                2       タイセイ
                3       コセイ
                4       リンセイ
5( 15-419) 2     ------ ハナ ノ タイショウメン ハ
                1       ホウシャカ
                2       ユウカ
6( 25-414) 3     ------ ハ ノ ヘリ ハ
                1       キレコム
                2       ノコギリハ
                3       センヘン
7( 27-92 ) 2     ------ フクガク ハ アルカ
                1       アル
                2       ナイ
8( 28-265) 4     ------ オシベノ カズ ハ
                1       タスウ
                2       10 ホン
                3       6 ホン
                4       2 ホン
9( 30-459) 2     ------ ヨウシュウ ハ アルカ
                1       アル
                2       ナイ
10( 31-460) 2    ------ クキ ハ チュウクウ カ
                1       ハイ
                2       イイエ
11( 32-118) 2    ------ クキ ニ ケ ハ アルカ
                1       アル
                2       ナイ
12( 33-417) 3    ------ ガクヘン ノ カズ ハ
                1       6 マイ
                2       4 マイ
                3       ナイ
```

図4. 質問の例

```
1     1 = 1 ( 2)
スミレ = マキヒゲ ハ アルカ? (ナイ)
2     1 = 2 ( 1)
スミレ = タクヨウ ハ アルカ? (アル)
3     1 = 3 ( 1)
スミレ = タンヨウ カ フクヨウ カ? (タンヨウ)
4     2 = 1 ( 2)
ヤエムグラ = マキヒゲ ハ アルカ? (ナイ)
5     2 = 3 ( 1)
ヤエムグラ = タンヨウ カ フクヨウ カ? (タンヨウ)
6     3 = 1 ( 2)
アメリカフウロ = マキヒゲ ハ アルカ? (ナイ)
7     3 = 2 ( 1)
アメリカフウロ = タンヨウ ハ アルカ? (アル)
8     3 = 3 ( 1)
アメリカフウロ = タンヨウ カ フクヨウ カ? (タンヨウ)
9     4 = 1 ( 2)
セニバアオイ= マキヒゲ ハ アルカ? (ナイ)
10    4 = 2 ( 1)
セニバアオイ = タンヨウ ハ フルカ? (アル)
11    4 = 3 ( 1)
セニバアオイ = タンヨウ カ フクヨウ カ? (タンヨウ)
12    1 = 4 ( 1)
スミレ = ハ ノ ツキカタ ハ? (コンセイ)
13    3 = 4 ( 2)
アメリカフウロ = ハ ノ ツキカタ ハ? (タイセイ)
14    4 = 4 ( 3)
セニバアオイ = ハ ノ ツキカタ ハ? (コセイ)
15    4 = 5 ( 1)
セニバアオイ = ハナ ノ タイショウメン ハ? (ホウシャカ)
16    5 = 1 ( 2)
タチスミレ = マキヒゲ ハ アルカ? (ナイ)
17    5 = 2 ( 1)
タチスミレ = タクヨウ ハ アルカ? (アル)
18    5 = 3 ( 1)
タチスミレ = タンヨウ カ フクヨウ カ? (タンヨウ)
19    5 = 4 ( 3)
タチスミレ = ハ ノ ツキカタ ハ? (コセイ)
20    5 = 5 ( 2)
タチスミレ = ハナ ノ タイショウ メン ハ? (サユウカ)
21    6 = 1 ( 2)
ハルタデ = マキヒゲ ハ アルカ? (ナイ)
22    6 = 2 ( 1)
ハルタデ = タクヨウ ハ アルカ? (アル)
```

図5. プロダクションルールの例

§8. あとがき

植物の色・形状などの特徴に基づいて，その植物名を決定する植物検索システムを試作した。このシステムはとくに知識の追加・訂正などを容易にするために，プロダクションルールの形で与えてある。しかし，検索する植物の種類が多く，したがって検索するための特徴が多い場合にはプロダクションルールによる手法では記憶容量が大きくなる。しかし，その半面欠損データや特種な特徴などがある場合とかあるいは知識の追加・訂正などの機能を有していることを考慮すれば，やはりテーブル方式よりプロダクションルールによる知識の表現形式の方が優れていると結論できよう。試作モデルでは約60種の植物についてであるが，今後さらに種類を増加するとともにまたそのデータ構造についても重要な課題として考察が必要であろう。

最後に，プログラミングおよび実験については，松田圭弘(YHP)，森光彦(ヤマハ)および石塚昭夫(院生)の三君の大きな協力があったことを付記しておく。

```
KENSAKU O HAJIMEMASU.
  32.83  26  54                                   TANYOU KA FUKUYOU KA?
  NI TSUITE TSUGINOUCHIKARA ERANDE KUDASAI.
       1-->              TANYOU
       2-->         5SHUTSU FUKUYOU
       3-->         3SHUTSU FUKUYOU
       4-->            UJOU FUKUYOU
03350 ?
4
   9.29  22  21                                   YUUSHOU WA ARUKA?
  NI TSUITE TSUGINOUCHIKARA ERANDE KUDASAI.
       1-->                 ARU
       2-->                 NAI
03350 ?
2
   6.44  19  16                                   JAKUYOU WA ARUKA?
  NI TSUITE TSUGINOUCHIKARA ERANDE KUDASAI.
       1-->                 ARU
       2-->                 NAI
03350 ?
2
   4.75  17  12                                   ZENTAI NO IRO WA?
  NI TSUITE TSUGINOUCHIKARA ERANDE KUDASAI.
       1-->           RYOKUSHOKU
       2-->          TANSHISHOKU
       3-->           TANOOSHOKU
03350 ?
1
   4.20  16  10                                   HANA NO TAISHOUMEN WA?
  NI TSUITE TSUGINOUCHIKARA ERANDE KUDASAI.
       1-->              HOUSHAKA
       2-->               SAYUUKA
03350 ?
2
   3.14   8   7                                   KI KI WA CHUUKUU KA?
  NI TSUITE TSUGINOUCHIKARA ERANDE KUDASAI.
       1-->                 YES
       2-->                  NO
03350 ?
2
   2.00   4   5                                   TOUJOUKA KA?
  NI TSUITE TSUGINOUCHIKARA ERANDE KUDASAI
       1-->                 YES
       2-->                  NO
03350 ?
EL
     KANOUSEI NO ARU SHOKUBU WA...
    PLANT                              KOUTEI KAISUU
     1              ITACHINOSASAGE          0
    31                KITSUNEAZAMI         5
    37                   KIKEMAN           6
    39                MURASAKIKEMAN        6
    40                SERIBAHIENSOU        6
      * IJOH  5 SHU *
   2.00   4   5                                   TOUJOUKA KA?
  NI TSUITE TSUGINOUCHIKARA ERANDE KUDASAI.
       1-->                 YES
       2-->                  NO
03350 ?
2
```

図 6.　検索の例

参考文献

1. 松田圭弘，“対話形式の植物検索システムの作成”，東京工大 情報工学科 卒業論文 1980.3
2. 森　光彦，“知識ベースを利用した情報検索システムの作成”，同上，1981.3
3. 石塚昭夫，“知識ベースと対話形式検索手法の研究”，同上，1982.3
4. 西村恕彦，“FORTRAN データベース”，bit 増刊号
5. 石戸　忠，“実践的植物検索小図鑑 1-2”，講談社，1979
6. 同　上，“学習コンピュータ植物図鑑”，文理書院
7. 長沢雅男，“情報検索入門”，森北出版，1971