

概念階層を形成する事例ベースの構造化

山田正隆 高田修 渡辺豊英

名古屋大学大学院工学研究科

〒464-01 愛知県名古屋市千種区不老町

Phone: 052-789-5143

E-mail: {yamada,takata,watanabe}@watanabe.nuie.nagoya-u.ac.jp

あらまし

事例ベース推論システムでは、事例ベースから新たな問題に類似する事例を効率よく検索するために、事例ベースの構造化について検討されてきた。そして、それらの多くは階層構造を指向し、個々の属性を構造の節点として扱ってきた。しかし、個々の属性より、属性集合の方が分類に対する寄与が大きいと考えられ、属性集合を事例構造に取り入れることは有効である。また、類似事例検索を実現するには、概念の階層的知識を用いた意味的な照合処理が有用であり、概念階層の属性を事例ベースの構造に取り入れることに大きな意義がある。本稿では、二つの事例に共通する特性に基づいて、個々の事例を階層化し、事例ベースを構成するアプローチを報告する。

キーワード

事例ベース推論, 事例ベースの構造, 類似検索

Case-Base Structure Incorporated with Attribute Set in Concept Hierarchy

Masataka YAMADA, Osamu TAKATA and Toyohide WATANABE

Graduate School of Engineering, Nagoya University

Furo-cho, Chikusa-ku, Nagoya 464-01, JAPAN

Phone: 052-789-5143

E-mail: {yamada,takata,watanabe}@watanabe.nuie.nagoya-u.ac.jp

Abstract

In Case-Based Reasoning(CBR) system, the case-bases have been ordinarily organized under the hierarchical structure so as to find out candidates effectively, which are similar or consistent to an input term. In order to categorize individual cases appropriately, characteristic attributes are assigned independently to different nodes in the hierarchical structure traditionally. However, sets of these attributes contribute to the categorization more effectively than a single attribute. In particular, it is powerful to apply sets of attributes to similarity search. In this paper, we discuss the approach to construct the hierarchical structure of case-base, based on the common properties between two cases.

Keywords

case-based reasoning, case-base structure, similarity search

1 はじめに

事例ベース推論 [1, 2] における重要な機能として類似事例検索がある。類似事例検索の場合、個々の事例を組織化しないフラットな事例ベース構造では、網羅的な探索が必要となり、膨大な探索時間を要する。そのため、一般に複数の事例間における特徴記述の共通性に着目して、カテゴリ化、分類木などを利用し、事例の汎化階層 (分類階層) を形成することにより、効率的な検索を図っている。

事例ベース推論では、弁別木のような階層的構造によって事例ベースを構成し、事例中の適当な特性をインデックスとして事例を探索する方法がよく用いられる。しかし、類似事例検索の場合には、一つの特性により事例を排他的に分類することは不適切である。類似検索の場合、多くの情報間の関係に基づいて類似性が判別されることが少なくなく、特定の特性により分類する方法は不適切である。複数の特性を階層構造の節点に持たせた方が、分類に対する貢献が大きくなり、より適切である。また、類似事例検索を実現するには、概念の階層的知識を用いた意味的な照合処理が有効であり、概念階層属性を事例ベース構造に取り入れることには大きな意義がある。

本稿では、複数の特性を階層構造の節点として設定し、概念階層属性を取り入れた事例ベース構造を構築するアプローチについて報告する。本アプローチでは、複数の特性を階層構造の節点として設定するため、二つの事例に共通する特性に基づいた構造化手法を採用する。この共通する特性に基づいた構造化手法では、構造の最下位から順に上位へ、それぞれの層で共通する同じ属性を抽出して、それを上位概念とし、再帰的に概念階層を構成する。

2 提案手法

提案する手法は、類似した二つの事例間で共通特性を作り、それを事例構造に組み入れることで、事例ベースを構造化する。まず、概念階層、扱う事例表現を述べ、さらに共通特性の構成法について述べた後、構造化のアルゴリズムについて説明する。

2.1 一般的定義

2.1.1 事例表現

個々の事例表現として、属性-値の組より成るタプルを採用し、属性値として概念階層属性を適用する。概念階層属性とは、離散値を取る属性であり、その属性値を is-a 階層からなる階層構造に配置することで、意味的に近い語ほど近くに配置されるという構造を持つ。例えば、属性「形」、「色」に対する概念階層はそれぞれ図1のような構造を持つ。

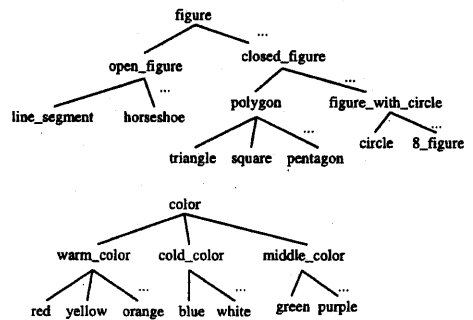


図1: 属性「形」、「色」に対する概念階層

そして、その属性値間に概念階層を適用することにより、個々の事例間に包含関係を導入し、また二つの事例間の共通特性に基づいて構造化する。

各事例 C は次のように表現される、ただし、 A_i は属性、 V_i は属性値を示している。

$$C = \langle \langle A_1, V_1 \rangle, \dots, \langle A_n, V_n \rangle \rangle$$

2.1.2 類似度計算

本構造化手法は類似度の計算方法に依存しないので、どのような計算方法を用いてもよいが、本稿では、二つの語の深さを d_i, d_j 、それらの最下位の共通上位概念の深さを d_c としたとき、属性値間の類似度を、

$$\text{similarity}(d_i, d_j) = \frac{d_c \times 2}{d_i + d_j}$$

とした [3]。例えば、図1の概念「triangle」と「square」の類似度を求める。それぞれの概念の深さは3、共通の上位概念「polygon」の深さは2な

ので、概念間の類似度は

$$\text{similarity}(\text{triangle}, \text{square}) = \frac{2 \times 2}{3 + 3} = \frac{2}{3}$$

となる。事例間の類似度は各属性値の類似度の重み付き和を用いた。ただし、簡単のため各重みを1としている。

2.2 概念階層と共通特性

事例ベース構造は、二つの類似した事例間において

- 1) 概念階層による共通の上位概念を構成
- 2) 共通の属性集合を構成

という二つの視点による共通特性により構成される。ここで、作成される共通特性は二つの事例を包含し、上位カテゴリと定義付けられる。このように構成することで、概念階層属性を構造に取り入れることができる。また、個々の属性より属性集合の方が分類における寄与は大きいので、より望ましい分類階層を形成することができる。

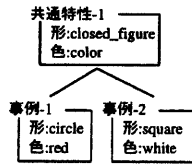


図 2: 共通特性

例えば、図 1 の概念階層において、図 2 の事例-1 と事例-2 の属性「形」に対する共通の上位概念は「closed_figure」、「色」に対して「color」となり、この上位概念が共通特性の持つ属性となり、図 2 のような共通特性を有し、構造化することができる。このとき、事例-1,2 はその共通特性に包含されている。

2.3 事例ベースの構築アルゴリズム

事例ベースの構築(新規事例を事例ベースに登録)はまず、新規に登録する事例に類似した事例を既存の事例ベースから一つ検索し、次に、それらの中で共通特性を作り、それを事例構造に組み入れる。事例構造を用いた事例の検索手法については後述する。

共通特性を事例構造に組み入れるため、共通特性を二つの事例間に設定し、次に新しく設定した共通特性の上位のリンク先を見つける。この上位のリンク先は共通特性を作成した事例の既存の上位節点によって異なるが、基本的にはその既存の上位節点と、新しく設定した共通特性との間でさらに、共通特性を作成する。このように、下位から上位へ、順に事例ベース構造を改編することで、構造を再構成し、事例ベースを逐次的に構築することができる。このアルゴリズムを図 3 に示す。

Top-level call: MakeTree(Searched-case, Input-case)
Results: updated structure

Procedure MakeTree(N, R)

N: current node of the structure

R: An instance to be registered

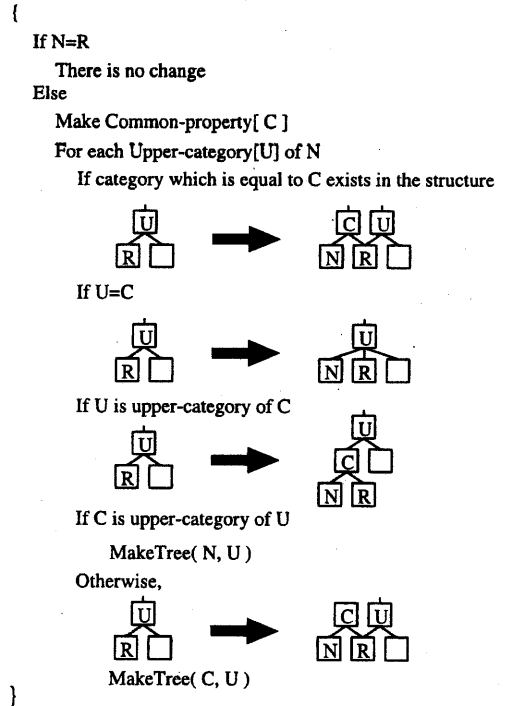


図 3: 事例ベース更新アルゴリズム

このアルゴリズムにより事例ベースを構築すると、図 4 の左上のような冗長なリンクが作成される場合がある。このようなリンクは検索性能の向上という点において必ずしも冗長とはいえないが、リン

クの数が増大になる恐れがあり、好ましくない。冗長なリンクができるのは図4に示す三つの場合で、新たにリンクが設定される時冗長なリンクの生成の有無を確認し、生成される場合にはそのリンクを削除する。

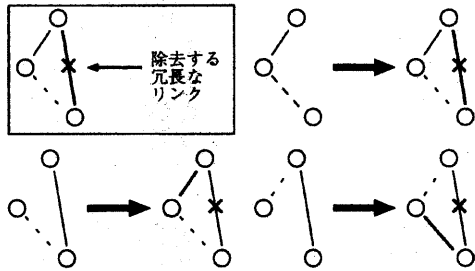


図4: 冗長なリンクの除去

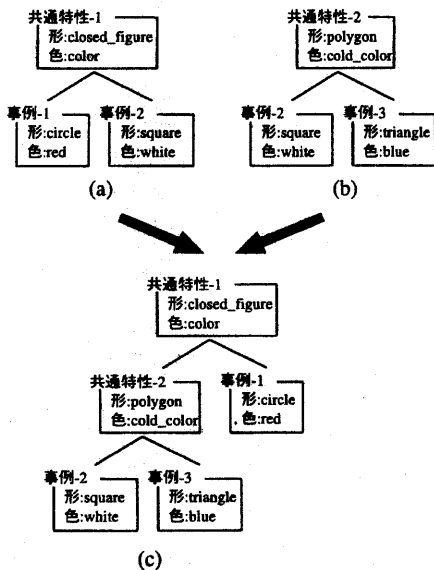


図5: 事例ベース構造の例-(1)

図5-(a)の事例構造に対して、事例-3が追加された場合を例として考える。まず、事例-3と既存の事例ベース構造中の類似する事例(事例-1)との間で共通特性を作成する(図5-(b))。次に作成した共通特性-2と事例-1の上位カテゴリ(共通特性-1)を比較する。この場合、共通特性-1は共通特性-2を包含しているため、共通特性-1を共通特性-2の上位に

配置し、図5-(c)のように構成することができる。

このようにして作成した事例ベース構造の例を図6に示す。図中の円は節点、四角は事例、数字は作成された順序を示している。

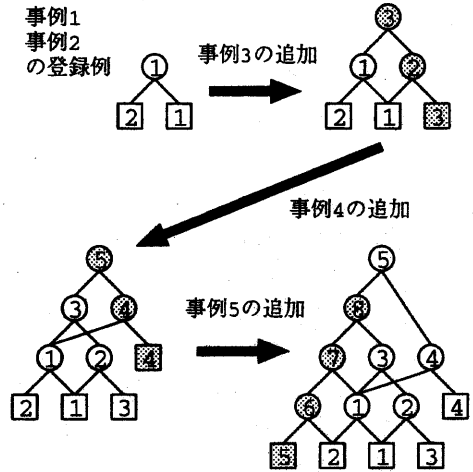


図6: 事例ベース構造の例-(2)

このアルゴリズムによって作成される事例ベース構造は、サイクルのない有向グラフとなる。一般に、事例ベース構造では、排他的な木構造として構成されるが、我々の手法では、複数の特性を構造に組み入れたことで、上位へのリンクを複数取るネットワーク構造となる。

2.4 類似事例検索

先のアルゴリズムにより作成した事例ベース構造を用いて類似事例を検索する方法について説明する。まず、検索事例を包含する上位カテゴリを全て検索する。事例構造は包含関係により上位下位が設定されており、上位カテゴリとなる節点は、その上位の節点もまた上位カテゴリとなっているので、事例構造に存在する上位カテゴリは構造のルートから包含する節点の一つずつたどることにより全て見つけることができる。

このようにして探索した上位カテゴリの中で、終端となる節点の下位節点を全て取り出し、それらと検索事例の間で類似度を計算し、最も類似度の高い節点を求める。例えば、図7の事例構造の場合、

まず上位カテゴリとして、黒丸の節点が探索されたら、次にそれらの中で終端となる節点の下位節点として、斜線の丸が類似度計算の対象として探索される。

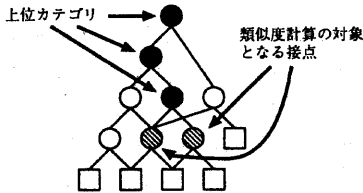


図 7: 事例検索の例

次にその節点の下位節点を全て取り出し、また検索事例との間で類似度を計算し、最も類似度の高い節点を求める。この手続きを事例構造の終端となる事例を探索するまで続けることで、類似事例を検索することができる。

3 実験

事例データを仮名漢字変換における Q&A とした 134 の事例を入力して実験した。この例では、属性値はさらに引数として、属性-値の組を取ることもでき、入れ子構造を取る。事例表現の例を図 8 に示す。

```
((VER (down)
  (AGT (jserver)
    (TIM () ((VER (connect))
      (AGT (egg))
      (GOL (jserver))
      (SIT (mule))))))
```

図 8: 事例表現の例

概念階層は Roget シソーラス [4] に手作業で不足する属性値を加えて用いた。

構造化を行わない場合との検索時間の比較結果を図 9 に示す。横軸は登録事例数、縦軸は検索時間の平均値を示している。さらに、この検索において探索された節点の数を図 10 に示す。図中の直線はフラットな構造のときに類似検索の対象となる事例の数を示している。

グラフを見て明らかのように、探索対象の数は構造化した場合の方が多くなっているのにもかかわらず、

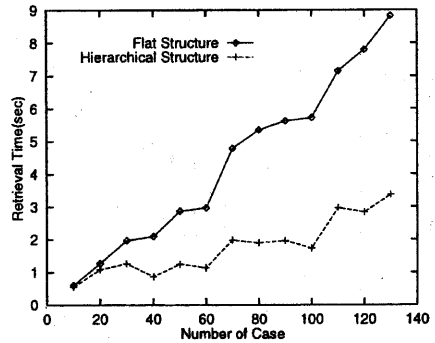


図 9: 検索時間の比較

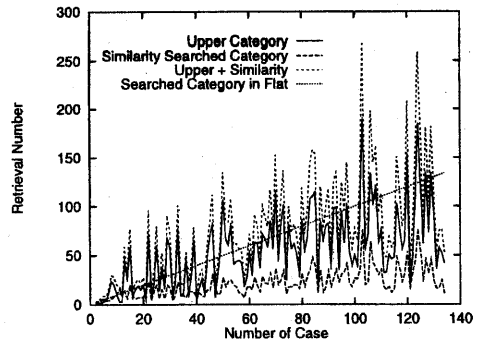


図 10: 探索節点

ならず、検索時間において改善が見られる。構造化した場合の探索は、その多くが検索事例の上位カテゴリとなっているかを確認することのみが行われ、類似度を計算している数は少ない。上位カテゴリとなるかの確認は、類似度計算にかかるコストより少ないこと、また中間節点が事例そのものよりも情報量が少ないため、比較にかかるコストが少ないことから、検索時間が改善されている。

構造化した場合、必ずしも最も類似した事例が検索される保証がないので、最後にこの探索において検索された事例と検索事例との間の類似度と、フラットな構造により検索された事例における類似度を比較し、分布図として図 11 に示す。133 回検索を行い、このうち最も類似した(フラットと同じ)事例が検索されたのは 67 回であった。

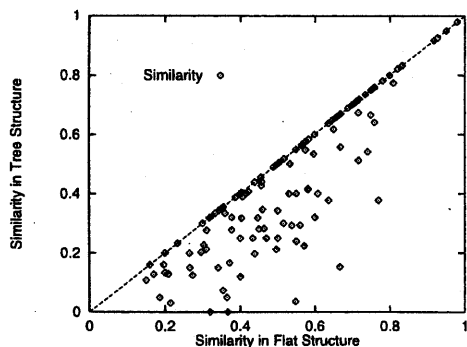


図 11: 分布

4 他の関連システムとの比較

概念階層属性を取り入れた事例検索法として、決定木学習を拡張した方法 [5] がある。しかし、この決定木学習の方法はクラス概念に従って分類するアプローチを採用していて、事例ベース推論における一般的な問題に適用することは困難である。

また、宇津呂 [6] により概念階層属性をとる事例に対する高速検索手法が提案されている。個々の事例とは独立に類似度計算のあらゆるパターンを表現した類似度計算テンプレートをを用いることで、厳密に最も類似した事例を事例数に依存せず、高速に検索することを実現している。

前提として、事例間の類似度計算の結果を適度な数の有限個のパターンに帰着できる必要があるとしているので、平衡な(ルートから終端概念までの階層の深さが一定である)概念階層に限られ、非平衡である概念階層の場合、概念間の類似度の数が増加するので対応できない。事例ベース推論では、状況として事例は逐次的に与えられるのが一般的であるので、概念階層もそれに伴い更新されると考えられる。この場合、葉の深さを変えずに概念階層を更新するのは困難であり、扱える状況も限定される。また、今回実験に用いたような、再帰的に表現される事例を扱うことができない。

5 おわりに

概念階層属性を事例ベース構造に取り入れる手法について述べた。これにより、意味的照合を考慮し

た事例ベース構造を形成することができる。

今後の課題としては、今回提案したアルゴリズムでは、不必要に多くの中間節点ができるので、必要以上に中間節点で作成されるのを防ぐため、中間節点を作成するときに、何らかの指標が必要と思われる。例えば、構造の深さを制限し、制限を越える場合はより特化されている情報量の多い節点を残すようにする、などが考えられる。また、本手法の評価には、実験に用いた事例数では不十分で、さらに多くの事例を登録して実験する必要がある。

謝辞

日頃、熱心に討論していただいている佐川雄二講師、朝倉宏一助手をはじめ、渡邊研究室の皆様感謝致します。

参考文献

- [1] J. Kolodner: "Case-Based Reasoning", Morgan Kaufmann (1993).
- [2] A. Aamodt and E. Plaza: "Case-Based Reasoning: Foundational Issues, Methodological Variation, and System Approaches", *AI Communications*, Vol. 7, pp. 39-59 (1994).
- [3] 長尾真: "自然言語処理", 岩波講座 ソフトウェア科学 15 (1996).
- [4] R. Chapman: "Roget's International Thesaurus (Fourth Edition)", Harper and Row, New York (1977).
- [5] フェインアルモアリム, 秋葉泰弘, 金田重郎: "木構造属性を許容する決定木学習", 人工知能学会誌, Vol. 12, No. 3, pp. 421-429 (1997).
- [6] 宇津呂武仁: "類似度計算テンプレートをを用いた検索質問生成による再近隣検索法の最適化", 人工知能学会誌, Vol. 11, No. 6, pp. 950-959 (1996).