2A2-1

形式的概念分析を用いた概念階層間の関係の発見

Discovery of Relationships among Concept Hierarchies using Formal Concept Analysis

市瀬 龍太郎[*1] 武田 英明[*1]
Ryutaro Ichise Hideaki Takeda

[*1]国立情報学研究所
National Institute of Informatics

One method by which to managing large amounts of information is to utilize catalogs which organize information within concept hierarchies. In the present paper, we address the problem of finding category relationships among multiple catalogs. The main idea of this paper is the use of formal concept analysis to find them. In order to evaluate the proposed method, we conducted an experiment. The obtained result shows that the proposed method successfully find out the relationships.

1. はじめに

大量の情報がある場合に，情報の整理法として概念階層を使った分類がしばしば用いられる．そのような例として，ファイルシステム，インターネットディレクトリ，商品カタログ，オントロジーなどが挙げられる．このような階層的に分類された情報が複数あった時には，一般的には異なる概念体系を使って分類されたものとなるため，分類情報をお互いに利用することが困難となる．本研究では，複数の概念階層を用いた情報分類があった時に，これらの分類階層のカテゴリ間の関係を自動的に発見する手法を提案する．本論文では，まずこの問題を概念階層間のカテゴリ関係発見問題として定式化する．次に，この問題に対応するために，概念的形式分析を用いた関係性発見手法を提案し，実験により，提案手法の有効性を検証する．

2. 概念階層間のカテゴリ関係発見問題

本研究で取り組む問題について，ここでは定式化を行う．問題で与えられるデータとして，下記のものを考える．

- 複数の概念階層
- インスタンスの集合とその概念階層内でのカテゴリ
- インスタンスの属性

1 点目の概念階層とは，情報をなんらかの形式で分類するときに使うものであり，階層性を持つものとする．2 点目のインスタンスとは，分類される情報自身を表す．例えば，ファイルシステムの場合は，ファイルが分類される情報のインスタンスに相当し，ディレクトリがカテゴリに相当する．3 点目のインスタンスの属性とは，インスタンスの性質を表すものである．例えば，ファイルをインスタンスとした時には，「ワードの文書である」や「○○について書かれた文書である」などが属性となる．ここでは，属性は 2 値を取るものとする．機械学習で使われる属性の形式では，多値を取るものも使われることが多いが，そのようなものは，複数の属性に分解し，2 値の属性にすることが可能である．従って，多値の属性を持つものに対しても，この問題の定式化は対応可能である．一方，ここで発見する関係は，それぞれの概念階層に含まれるカテゴリ間の関係となる．

連絡先: 市瀬 龍太郎，国立情報学研究所情報学プリンシプル研究系，〒 101-8430 東京都千代田区一ツ橋 2-1-2，Tel:03-4212-2000，Fax:03-3556-1916，E-mail:ichise@nii.ac.jp

3. 形式的概念分析を用いた関係の発見

関係の発見を行うために，まず，人間の作る概念階層について考察をしてみる．人間が分類のために作る概念階層として，2 種類の概念階層が考えられる．ここでは，それを，決定木と分類木と呼ぶことにする．決定木は，機械学習でよく使われている C4.5 [Quinlan 93] でも用いられているものであり，インスタンスを概念階層の最上位から評価し，決定的に分類する方法である．各々の階層では，特定の属性を評価し，その評価に基づいて分類先を決める方法である．この概念階層では，決定的に分類されるため，分類結果が分かり易いという特徴がある．一方，分類木は，インスタンスを最上位の階層から評価，分類する点では，決定木と同じであるが，分類の際に，ある属性を持つということに基づいて並列に分類を行う点で，決定木と異なる．例えば，決定木では，ある階層において，属性 A で分類すると，A の値によって，その下の階層の行き先が決まるのに対して，分類木では，ある階層において，属性 A を持つならば，その下の階層に分類され，同時に，属性 B を持つならば，そちらの階層の下にも分類されるというように，並列で評価が行われる．結果として，分類木においては，インスタンスは複数のカテゴリに分類される可能性を持つことになる．

では，人間の情報分類は，決定木と分類木のどちらを通常用いているのであろうか．人間は，ある情報について，全ての属性を知っているわけではないし，全ての属性の情報をいつも確定させているとも限らない．そこで，分類の際には，決定木よりも分類木のような緩い分類手法を用いていると考えられる．事実，キーワードを用いた文書の分類などでは，分類木の手法が用いられているし，自動車の運転制御の規則では，人間の状況分類が分類木と親和性を持っていることが示されている [市瀬 04]．

そこで，本研究では，可能性のある分類木を列挙することができる形式的概念分析 (Formal Concept Analysis) [Ganter 99] を用いて階層間の関係の発見を試みる．形式的概念分析とは，属性の関係を用いて概念束を構成する手法である．例えば，表 1 のようなインスタンスと属性の例があった時に，図 1 のような概念束を導出する．

本研究では，この概念束を用いて関係性の発見を行う．まず，それぞれの概念階層に対して，形式的概念分析を実施し，それぞれのカテゴリが形式的概念分析の結果とどのように対応するかを計算する．次に，全てのインスタンスを用いて，形式的概念分析を行う．その結果，それぞれの階層におけるカテゴリが全体として見た時に，概念束のどこにマッピングされるこ

表 1: インスタンスと属性の関係の例

	水中で呼吸 A	飛行 B	手 C	翼 D
こうもり		○		○
さる			○	
ペンギン				○
さめ	○			

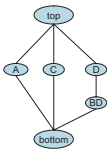

図 1: 形式的概念分析で作られた概念束の例

とになるのかが分かる．このとき，概念束の中で，それぞれのカテゴリがどのような関係になるのかが分かるため，結果として各概念階層におけるカテゴリがどのような関係になるのかということを同定することが可能となる．

4. 実験

本研究で提案した手法の性質を調べるために，実験を行った．実験では，8 属性を組み合わせた 128 個のインスタンスを人工的に生成し，全インスタンスの中からランダムに 32 個を抽出してインスタンスセットを 2 つ生成した．その 2 つのインスタンスセットに対して，あらかじめ規定した 2 つの概念階層で分類を行い，それを実験データとして使用した．図 2 に，実験に使った概念階層の一つを示す．図 3 は，図 2 の概念階層のデータから形式的概念分析を用いて作られた概念束である．図中の同じ色で塗られた部分は，同じカテゴリに属するインスタンスのみが集まった概念である．図の色は図 2 と図 3 で対応しており，同じカテゴリを表している．次に，図 4 に全部のインスタンスデータを用いた形式的概念分析の結果を示す．

図 4 から分かるように，全体のインスタンスのデータを用いて構築した形式的概念分析では，2 つの概念階層の中にあるカテゴリが出現しており，兄弟関係にあるカテゴリも示されている．このように，提案手法を用いると，2 つの概念階層におけるカテゴリ間の関係を同定することが可能であることが示された．一方，形式的概念分析では，属性が存在していることのみを使って，概念束を構成するため，属性が不在であるという情報の取り扱いが難しいことが分かる．これは，表 1 の例において，「飛ばない」という情報 ($\neg B$) を使ってカテゴリの分類がなされているような場合には，それらの情報が概念束に出てこないため，関係性を取り出すことができない．このような問題に対処するためには，新たに $\neg B$ という情報を属性に付加してから概念束を構築することで，対処が可能であると考えられる．

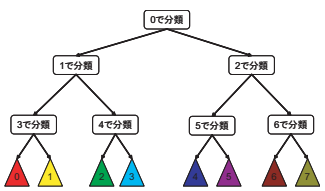

図 2: 実験に使った概念階層の一つ

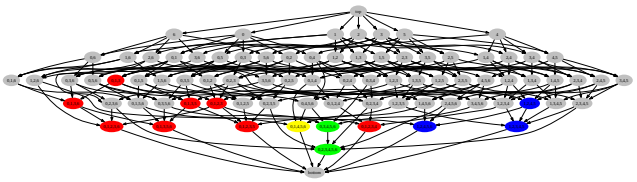

図 3: 概念階層から作られた概念束

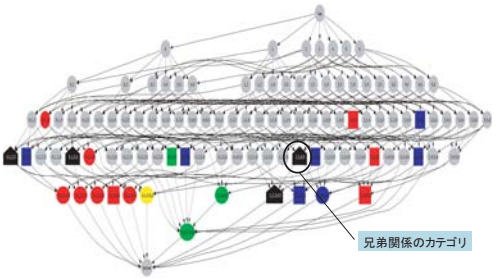

図 4: 全てのインスタンスから作られた概念束

5. むすび

本研究では，概念階層間の関係を抽出するという問題に対して，新たに形式的概念分析を用いることで，異なる概念階層のカテゴリ間の関係を抽出する方法を提案した．そして，その提案手法を用いて実験を行い，カテゴリ間の関係が抽出できることを示した．同様のアプローチを取った研究に，Stummeらの FCA-Merge [Stumme 01] がある．FCA-Merge では，2 つのオントロジーを統合した新たなオントロジーを生成するのに形式的概念分析を用いているが，本研究では形式的概念分析を用いることで，2 つのオントロジーのカテゴリ間の関係の同定を行っている．今後の課題としては，より広範なデータに関して実験を行い，本手法の性質を明らかにすると同時に，多数の概念束が出てきた時に，効率よく関係を列挙する手法の確立が必要となる．また，オントロジーマッチングなどの実データを使った検証も行っていく必要性がある．

参考文献

[Ganter 99] Ganter, B. and Wille, R.: Formal Concept Analysis: Mathematical Foundations, Springer, (1999).

[市瀬 04] 市瀬 龍太郎，ダニエル シャピロ，パット ラングリー: 行動履歴からの構造的プログラムの学習法，電子情報通信学会論文誌 D-1，Vol. J87-D-1, No. 6, pp. 730-740, (2004).

[Quinlan 93] Quinlan, J. R.: C4.5: Programs for Machine Learning, Morgan Kaufmann, (1993).

[Stumme 01] Stumme, G. and Maedche, A.: FCA-Merge: Bottom-Up Merging of Ontologies, In Proceedings of the 17th International Joint Conference on Artificial Intelligence, pp. 225–230, (2001).