Vol.2009-MPS-76 No.14
Vol.2009-BIO-19 No.14
2009/12/17

萌芽的閉包を枚挙する分枝限定法について

中島健志[†1]　原口誠[†1]　大久保好章[†1]

本稿では，低頻度の形式概念抽出問題について議論する．顕在パターンマイニング・コントラストセットマイニング・クロスオーバ概念抽出等の研究に見られる通り，レア概念・低頻度パターンの重要性が認識されている一方で，概念の意味解釈の困難さの問題が同時に存在する．すなわち，外延の小さなレア概念は，一般に多くの属性から成る内包を伴うことから，その明確な解釈を与えることが容易ではない．

こうした問題に対処すべく，本稿では，抽出すべきレア概念を『少数の一般的な属性で特徴付けられる概念の解釈・理解は容易である』との考えのもとに定め，それらを**簡潔なレア形式概念**として定式化する．さらに，分枝限定法に基づくレア形式概念抽出アルゴリズムを提案し，計算機実験により，容易に解釈可能なレア概念が抽出できることを確認する．

A Branch-and-Bound Algorithm for Extracting Concice and Rare Concepts

Takeshi NAKAJIMA , Makoto HARAGUCHI and Yoshiaki OKUBO

In this paper, we present an algorithm for finding concepts (closures) with smaller supports. As suggested by the study of emerging patterns, contrast sets or crossover concepts, we regard less frequent and rare concepts.

However, we have several difficulties when we try to find rare concepts. In general, te lengths of rare concepts become longer, involving many attributes at various levels of generality. Consequently, it becomes harder to understand what the concepts mean.

In order to solve the above problems, we make a restriction about formation processes of concepts, where the formation is a flow of adding attributes to the present concepts already formed. The present concepts work as conditions for several candidate attributes to be added to them. Given such a present concept, we prohibit adding attributes strongly correlated with the present concept. As a result, the detected concepts has lower supports and consist of only attributes directing at more specific concepts through the formation processes.

We design an algorithm is designed as a top-N closure enumerator using branch-and-bound pruning rules so that it can reach concepts with lower supports by avoiding useless combination of correlated attributes in a huge space of concepts. We experimentally show the effectiveness of algorithm and the conceptual clarity of detected concepts because of their shorter length in spite of their lower supports.

1. はじめに

データマイニング研究の主要なテーマのひとつとして，**飽和アイテム集合**，あるいは，それと等価な**形式概念**の抽出・列挙問題が注目されて久しい．これらの研究では主に，生起頻度が比較的大きい**頻出パターン**が抽出のターゲットとされてきた[2),8)]．頻出するものは重要であるとの認識は，重要度を測る上でのひとつの有効な経験則であり，一般にも広く受け入れられよう．しかし，一方で，非頻出なものの中にも重要なものが潜む可能性もあり，これらを抽出のターゲットとする立場も重要である．

例えば，**顕在パターンマイニング** (*emerging pattern mining*) [3),4),6)] や**コントラストセットマイニング**[5),6)] では，ある条件下で非頻出な**レアパターン** (*rare pattern*) に注目している．具体的には，あるデータベース DB_1 では頻度の低いパターンが，もう一方のデータベース DB_2 ではより高い頻度を示す時，そのパターンは何らかの重要な変化を示唆すると考え，それを顕在パターンと呼び，抽出のターゲットとしている．

これに対して，著者等は文献 12) において，形式概念解析の枠組に基づき，概念 A と B それぞれの特殊概念の共通の汎化概念に相当する X を，A と B の**クロスオーバ概念**と呼び，その抽出を試みた．こうした X は，例え外延が小さなレアな概念であったとしても，メジャーな概念 (ここでの A と B) 間の，通常とは異なる視点からの隠れた継りを示唆するものと考えられ，さらに精査する価値があると期待できる．

こうした非頻出な概念・パターンの重要性に鑑み，本稿でも，小さな外延を有するレア概念の抽出に焦点を当てる．しかし，形式概念の理論的性質より，小さな外延を有する概念は，一般に大きな内包を伴う．例えば，著者らがこれまでの研究で行なった，1,223 の語彙で表現された 11,000 文書のデータに対する実験では，わずか 11 文書から成る小さな外延に対し，その内包を構成する語彙数は 111 となる形式概念が得られている．しかし，容易

†1 北海道大学大学院情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University

Vol.2009-MPS-76 No.14
Vol.2009-BIO-19 No.14
2009/12/17

に想像できる通り，100 を超える語彙集合をもとに，それら文書群が定める概念の意味を捉えることは極めて困難であることは言うまでもない．

こうした問題に対処すべく，本稿では，抽出すべきレア概念を

【少数の一般的な属性により特徴付けられる概念の意味解釈・理解は容易である】

との考えに基づいて定め，それらを**簡潔なレア形式概念** (*Concise and Rare Concepts*) として定式化し，その抽出アルゴリズムを与える．そこでは特に，少数属性による特徴付けを可能とするために，**属性の追加過程に伴う外延の変化に注目**し，また，一般的な属性による特徴付けを可能とするために，属性頻度に基づく一般性を測る評価尺度を導入し，その最大化問題を解くことで，これを実現する．

2. 準　　備

形式概念解析 (*Formal Concept Analysis*)[1] は，個体集合間の意味的な構造を解析する枠組のひとつである．

個体 (*object*) の集合 G，および，**属性** (*attribute*) の集合 M に対して，関係 $I \subseteq G \times M$ を考える．この時，タプル (G, M, I) を**形式文脈** (*Formal Context*) と呼ぶ．$(g, m) \in I$ の時，個体 g は属性 m を有すると言う．個体 g が有する属性の集合 $\{m \in \mathcal{M} \mid (g, m) \in I\}$ を，$M(g)$ で参照する．

形式文脈 (G, M, I) に関して，写像 $\varphi : 2^G \to 2^M$ および $\psi : 2^M \to 2^G$ を考える．ここで，個体集合 $X \subseteq \mathcal{G}$ と属性集合 $Y \subseteq \mathcal{M}$ について，

$$\varphi(X) = \{m \in M \mid \forall g \in X, \quad (g, m) \in I\},$$
$$\psi(Y) = \{g \in G \mid \forall m \in Y, \quad (g, m) \in I\}$$

とする．つまり，φ は X 中のすべての個体が共有する属性集合を，一方，ψ は Y 中のすべての属性を有する個体集合を返す写像である．

これら写像のもと，個体集合 $X \subseteq G$ と属性集合 $Y \subseteq M$ について，$\varphi(X) = Y$ かつ $\psi(Y) = X$ が成り立つ時，X と Y の組 $FC = (X, Y)$ を**形式概念** (*Formal Concept*) [1] と定める．ここで，X と Y をそれぞれ FC の**外延** (*extent*)，および，**内包** (*intent*) と呼ぶ．φ と ψ の定義より，$\psi(\varphi(X)) = X$ かつ $\varphi(\psi(Y)) = Y$ であることは明らかである．すなわち，形式概念とは，写像 φ と ψ に関して**閉じた** (*closed*) 個体集合 X と属性集合 Y の組で与えられる．X は，Y 中のすべての属性を有する個体のみから成り，かつ，それら以外にこうした個体は存在しない．同様に，Y は，X 中のすべての個体に含まれる (共有される) 属性のみから成り，かつ，それら以外にこうした属性は存在しない．

形式概念 $FC = (X, Y)$ および $FC' = (X', Y')$ について，$X \subseteq X'$ $(Y \supseteq Y')$ である時，かつ，その時に限り FC と FC' 間に順序関係を定め，これを $FC \preceq FC'$ と表記する．この時，FC は FC' の**特殊概念**, 逆に，FC' は FC の**汎化概念**と呼ぶ．所与の形式文脈におけるすべての形式概念の集合を $\mathcal{FC}$ とすると，順序関係 $\preceq$ のもと，$(\mathcal{FC}, \preceq)$ は束を構成し，これを**形式概念束** (*Formal Concept Lattice*) と呼ぶ．

形式文脈 (G, M, I) における個体と属性間の二項関係 I は，パターンマイニング研究における，トランザクションデータに他ならない．つまり，個体をトランザクション，属性をアイテムと考えれば，形式概念 (X, Y) の内包 Y は**飽和アイテム集合** (*closed itemset*)[8] に，また，外延 X は (飽和) アイテム集合 Y を含むトランザクション集合に対応する．よって，外延 X の大きさ $|X|$ は，アイテム集合 Y の**頻度** $sup(Y)$ と一致する．以降では，$|X|$ の代わりに $sup(Y)$ を用いることもある．

3. 簡潔な内包生成元を有するレア形式概念

先に述べた通り，本稿では，外延が比較的小さな形式概念，すなわち，レア形式概念の抽出を試みるが，形式概念の性質から，外延が小さくなるにつれ，その内包は次第に大きくなる．抽出された概念の意味は，その内包を構成する属性群に基づいて解釈することができるが，言うまでもなく，この様な大きな内包に対して適切な意味解釈を与えることは容易ではない．特に，内包が特殊 (specific) な属性ばかりを含む場合，その解釈はより一層困難なものとなることから，この様な概念は実際にはほとんど役に立たず，むしろ積極的に排除すべきであると考えられる．

そこで本稿では，レア概念の意味解釈を明確かつ容易にすべく，抽出すべき概念は，より一般的 (general) な少数の属性から成る内包を有するべきであるとの立場に立つ．以下，こうしたレア概念抽出問題の定式化，および，その抽出アルゴリズムについて議論する．

まず，形式概念の**レアネス** (*rareness*) を次の通り定義する．

定義 3.1　(形式概念のレアネス)
$FC = (X, Y)$ を形式概念，R を**レアネス閾値**とする．FC の外延のサイズ，すなわち，$|X| = sup(X)$ が R 以下である時，FC を R-**レア形式概念**と呼ぶ．■

形式概念 $FC = (X, Y)$ について，$\psi(Y) = X$ なる関係が成り立つことから，Y は X を特徴付ける属性集合であると考えることができるが，一般には，$\psi(Y') = X$ となる Y の

Vol.2009-MPS-76 No.14
Vol.2009-BIO-19 No.14
2009/12/17

部分集合 Y' が存在する．つまり，X は Y 中の一部の属性のみで完全に特徴付けることが可能である．この意味で，Y' はもとの Y と等価であり，FC のより簡潔な解釈を与え得ることから，本稿では，これを Y の**生成元** (*generator*) と呼び，ひとつの概念定義と見做し特に重視する．

定義 3.2 (内包の生成元)
$FC = (X, Y)$ を形式概念とする．$Y' \subseteq Y$ なる Y' について，$\psi(Y') = X$ である時，Y' を Y の**生成元** (*generator*) と呼ぶ． ■

ここで，内包 Y の生成元は一般に複数存在することに注意する．上述した通り，生成元は概念の意味解釈を与える重要な手掛かりとなることから，少数の一般的な属性から成る簡潔なものであることが期待される．この様な簡潔な生成元を有し，かつ，レアである形式概念を抽出するために，ここでは，生成元に対してある制約を課す．

形式概念 (X, Y) について，Y のある生成元 $Y' = \{m_1, \ldots, m_k\}$ を考える．$\psi(Y') = X$ より，外延 X は Y' により一意に同定されるが，ここでは，X が**生成元により同定される過程**，すなわち，Y' 中の属性を順次考慮していくことで，X が段階的に同定される過程に注目する．いま，Y' 中の属性 m_i に対して，$m_1 \prec \cdots \prec m_k$ なる順序が仮定されているとする．まず，一番目の属性 m_1 により，個体集合 $\psi(\{m_1\})$ が定まる．次に，属性 m_2 を追加すると，同定される個体集合は，$\psi(\{m_1\})$ から $\psi(\{m_1, m_2\})$ に縮小される．順次こうした処理を繰り返すと，最終的には次の個体集合の系列が得られる．

$$\psi(\{m_1\}) \supseteq \psi(\{m_1, m_2\}) \supseteq \cdots \supseteq, \psi(\{m_1, m_2 \ldots, m_{k-1}\}) \supseteq \psi(\{m_1, m_2 \ldots, m_k\}) = X$$

この様に，最初の個体集合 $\{b_1\}'$ が段階的に縮小されて X に辿り着くが，この X に至る過程が，概念を解釈・理解する際の容易さに大きな影響を与えるものと著者等は考える．このことをより詳細に述べよう．

隣り合う個体集合 $\psi(\{m_1, \ldots, m_{i-1}\})$ および $\psi(\{m_1, \ldots, m_{i-1}, m_i\})$ において，もし，これらがほぼ同じものであるならば，属性 m_i が，X を同定するにあたって顕著な役割を果たしているとは言えない．つまり，m_i の有無が X の同定過程に与える影響はほんのわずかである．X を同定可能な属性群が少ない程，形式概念 (X, Y) の意味理解が容易となることから，この様な m_i は**冗長** (*redundant*) なものであると考えられる．別の言い方をすると，任意の隣り合う個体集合間に，ある度合以上の縮小の様子が観測されるならば，その生成元は概念の容易な解釈・理解に大きく貢献するものと期待できよう．生成元に対するこうした要請は，次の通り定式化できる．

定義 3.3 (内包の δ-生成元)
(X, Y) を形式概念，$Y' = \{m_1, \ldots, b_m\}$ を Y の生成元とし，$m_1 \prec \cdots \prec m_k$ なる順序を仮定する．いま，**最小落差閾値** δ $(0 \leq \delta < 1)$ のもとで，任意の $m_i \in Y'$ $(1 \leq i < k)$ について，

$$\frac{sup(\{m_1, \ldots, m_{i+1}\})}{sup(\{m_1, \ldots, m_i\})} \leq 1 - \delta$$

ならば，Y' を Y の δ-**生成元**と呼ぶ． ■

Y' が Y の δ-生成元ならば，仮定された順序に従って Y' 中の属性を追加していく過程で，各段階で同定される個体集合のサイズは，少なくとも $(100 \times \delta)$ % ずつ縮小される．すなわち，生成元のサイズはパラメータ δ によって制御される．δ の値を大きくすることで，よりコンパクトな δ-生成元が得られる．ただし，パラメータ設定によっては，δ-生成元を持たない形式概念も存在する．

上記定義において，生成元中の属性間に仮定される順序は，生成元の簡潔さを考える上で極めて重要となる．生成元が m の属性から成る場合，仮定可能な順序は $m!$ 通り存在する．しかし，例えば，$\{m_1 \preceq m_2 \preceq m_3\}$ が δ- 生成元であったとしても，$\{a_1, a_3, a_2\}$ は必ずしもそうはならない．よって，仮定可能な順序の中から適当な順序を選択する必要があるが，本稿では，著者等の直感に従って妥当と思われる次の順序を仮定する．より具体的には，外延を特徴付ける過程で，より一般的なものから順次属性を追加した方が，その意味をより明確に理解できるであろうとの直感による．すなわち，**一般的なものから具体的なものへと順序付けられた属性群から成る生成元は，解釈・理解が容易である**との経験則に従い，ここでは，$sup(m_1) \geq \cdots \geq sup(m_k)$ なる順序を満たす生成元 $\{m_1, \ldots, m_k\}$ のみを妥当なものと考える．以下の議論では，δ-生成元は必ずこの要請を満たすものと仮定する．

これに加え，本稿では，生成元に**一般性**の尺度を導入する．解釈理解が容易な生成元は簡潔であるものが望ましいが，簡潔性には，それを構成する属性の一般性が深く関係する．一般的な属性から構成される生成元と，具体的な属性から成るそれとを比較すると，意味解釈の容易さからは，前者の方が望ましいことは経験的に明らかであろう．これに照らして，ここでは属性集合の一般性を次の通り定義する．

Vol.2009-MPS-76 No.14
Vol.2009-BIO-19 No.14
2009/12/17

定義 3.4 (属性集合の一般性)
(G, M, I) を形式文脈とする．属性集合 $Y \subseteq M$ の**一般性** (*generality*) を，それを構成する属性の頻度の最小値で定義し，これを $generality(Y)$ で参照する．すなわち，

$$generality(Y) = \min_{m \in Y}\{sup(m)\}$$

とする． ■

定義より，一般性の高い生成元は高頻度の属性のみから構成されることから，一般性の高いコンパクトな生成元の解釈は比較的容易であることが期待できる．また，任意の属性集合 $Y, Y' \subseteq M$ について，$Y \subseteq Y'$ ならば $generality(Y) \geq genarlity(Y')$ が成立する．つまり，一般性の値は，属性集合の拡張に伴い**単調減少**することに注意する．

以上の議論を踏まえ，本稿で扱う簡潔なレア形式概念抽出問題を次の通り定義する．

定義 3.5 (簡潔なレア形式概念の Top-N 枚挙問題)
(G, M, I) を形式文脈，δ を最小落差閾値，R をレアネス閾値とする．この時，(G, M, I) のもとで，以下の条件を満足する極大な形式概念 (X, Y) を抽出する問題を，**簡潔な** R**-レア形式概念の** *Top-N* **枚挙問題**と呼ぶ．

レアネス (制約)：
(X, Y) は R-レア形式概念，すなわち，$|X| = sup(A) \leq R$ である．

一般性 (目的関数)：
内包 Y の δ-生成元 Y' の一般性，すなわち，$generality(Y')$ の値は，任意の R-レア形式概念のそれらの中で上位 N 以内である． ■

次章では，簡潔なレア形式概念の Top-N 枚挙アルゴリズムについて議論する．

4. 簡潔なレア形式概念の Top-N 枚挙アルゴリズム

本章では，所与の形式文脈 (G, M, I)，最小落差閾値 δ，および，レアネス閾値 R のもとで，δ-生成元の一般性が上位 N 以内である極大な R-レア形式概念を抽出する**深さ優先分枝限定アルゴリズム**を与える．本アルゴリズムは，富田等による一連の最大クリーク抽出アルゴリズム[9),10)] の拡張として設計された Top-N 形式概念抽出アルゴリズム[12)–15)] を基礎としている．

4.1 基本探索戦略

(G, M, I) における任意の形式概念 (X, Y) について，$\psi(Y') = X$ かつ $\varphi(\psi(Y')) = Y$ となる $Y' \subseteq M$ が必ず存在する．よって，M の任意の部分集合に写像 ψ および φ を適用することで，(G, M, I) におけるすべての形式概念を枚挙することが可能となる．

いま，$M = \{m_1, \ldots, m_{|M|}\}$ 上にある全順序 $m_1 \prec \cdots \prec m_{|M|}$ を仮定し，M の部分集合 Y 中の属性はこの順序に従ってソーティングされているものとする．また，M の部分集合 $Y_i = \{m_{i_1}, \ldots, m_{i_k}\}$ について，Y_i の最終要素 m_{i_k} を $tail(Y_i)$ で参照する．さらに，先頭の ℓ 要素の集合 $\{m_{i_1}, \ldots, m_{i_\ell}\}$ を，Y_i の ℓ-接頭辞と呼び，$prefix(Y_i, \ell)$ で参照する．特に，$prefix(Y_i, 0) = \phi$ とする．

ここで，次の通り定義される 2^M 上の半順序 $\prec_s$ を導入する．

定義 4.1 (2^M 上の半順序)
M の部分集合 Y_i および Y_j を考える．Y_i が Y_j の $|Y_i|$-接頭辞である，すなわち，$Y_i = prefix(Y_j, |Y_i|)$ である時，かつ，その時に限り Y_i は Y_j の**前者** (*predecessor*) であると言い，これを $Y_i \prec_s Y_j$ と表わす．また，B_i が B_j の直接の前者である時，B_j を B_i の**子供** (*child*) と呼ぶ． ■

半順序集合 $(2^M, \prec_s)$ は，空集合 ϕ をルートノードとする**木** (*tree*) を構成し，これを**集合列挙木** (*set enumeration tree*) と呼ぶ．

集合列挙木中の各内部ノードに対応する $Y \subseteq M$ について，その子供を得るには，単に，$tail(Y) \prec m$ なる任意の属性 m を用いて $Y \cup \{m\}$ とすればよい．この様に，空集合を起点にこうした処理を繰り返すことで，M のすべての部分集合を，**機械的に，かつ，重複無く**列挙することができる．その際，後に述べる枝刈りが利用できることから，本稿では特に**深さ優先**で集合列挙木を探索するものとする．

集合列挙木を利用することで，すべての形式概念を容易に抽出することができる．具体的に述べると，各部分集合 $Y \subseteq M$ について，外延 $\psi(Y)$，および，内包 $\varphi(\psi(Y))$ を計算することで，形式概念 $(\psi(Y), \varphi(\psi(Y)))$ を抽出することができる．ここで，Y は内包 $\varphi(\psi(Y))$ の生成元となることから，集合列挙木により，可能な生成元をすべて調べることが可能となる．特に，M 中の属性に対して，その頻度 (sup 値) 降順の順序を仮定することで，前章の議論で要請した，一般的な属性から具体的なものへと並んだ生成元を調べることができる．

Vol.2009-MPS-76 No.14
Vol.2009-BIO-19 No.14
2009/12/17

[⋆1].

上記要請を満たすものの中で，本稿では特に，δ-生成元に興味があることは先に述べた．もし，ある生成元 Y が δ-生成元でないならば，$Y \prec_s Z$ なる任意の生成元 Z もまた δ-生成元ではない．よって，この様な場合には，Y の任意の後者 (子孫) を調べる必要はなく，直ちにバックトラックすればよい．

各形式概念の内包は，一般に複数の生成元を有する．すなわち，集合列挙木中の異なる生成元から同一の形式概念が抽出されてしまう．形式概念の効率の良い抽出を行なうためには，こうした重複生成の扱いが極めて重要となる．これに関しては後に議論する．

形式概念の理論的性質より，$Y \subset Z$ なる生成元 Y と Z から抽出されるそれぞれの形式概念 $(\psi(Y), \varphi(\psi(Y))$ と $(\psi(Z), \varphi(\psi(Z))$ の間には，形式概念束中で $(\psi(Z), \varphi(\psi(Z)) \preceq (\psi(Y), \varphi(\psi(Y))$ なる関係がある．いま，ここでのターゲットは，極大なレア形式概念の中で，その内包の生成元の一般性が上位 N 以内のものである．よって，集合列挙木の深さ優先探索において，ある生成元 $Y \subseteq M$ からレア形式概念が抽出された時点で，Y の任意の後者を調べる必要性がなくなることから，直ちにバックトラックが可能となる．

以上の議論より，簡潔なレア形式概念の Top-N 抽出のための基本探索戦略は次の通りとなる．空集合 ϕ を起点とし，順序関係 $\prec_s$ のもとで，各生成元を深さ優先で調べていく．探索の過程では，それまでに見つかった上位 N の簡潔なレア形式概念を保持するリスト，すなわち，**暫定的な** Top-N リストを管理する．生成元 $Y \subseteq M$ について，最初に Y が δ-生成元であるかを調べる．もしそうであれば，対応する形式概念 $(\psi(Y), \varphi(\psi(Y))$ を計算し，そのレアネスを調べる．もしレア形式概念であれば，暫定 Top-N リストを適切に更新する．つまり，δ-生成元 Y の一般性が，暫定リスト中 N 位のそれよりも高い，あるいは，等しい場合は，$(\psi(Y), \varphi(\psi(Y))$ を新規にリストへ登録し，バックトラックする．もし暫定 N 位に及ばない場合は，そのまま棄却し，バックトラックする．$(\psi(Y), \varphi(\psi(Y))$ がレア概念でない場合は，Y の子供を生成し，それに対して同様の処理を再帰的に繰り返す．こうした処理を，調べるべき生成元がなくなるまで繰り返せばよい．

4.2 不要な探索枝の枝刈り

簡潔なレア形式概念の Top-N 抽出を効率的に行なうためには，ターゲットに至らない不要な生成元を，探索の対象から積極的に除外する必要がある．ここでは，不要な探索枝をカットする枝刈り機構について述べる．

4.2.1 非 δ-生成元の排除

先に述べた通り，生成元 $Y \subseteq M$ に対して，その子供は，$tail(Y) \prec m$ なる属性 $m \in M$ を用いて $Y \cup \{m\}$ として得られる．こうした m をここでは Y の**拡張可能候補**と呼ぶ．また，Y の拡張可能候補の集合を $cand(Y)$ で参照する．すなわち，$cand(Y) = \{m \in M \mid tail(Y) \prec m\}$ である．

$cand(Y)$ 中の各属性に対して Y からの探索枝が張られることから，見込みのない候補属性を $cand(Y)$ から削除することで，不要な探索を枝刈ることができる．ここでの探索では，δ-生成元のみが必要なことから，属性 $m \in cand(Y)$ について，

$$\frac{sup(Y \cup \{m\})}{sup(Y)} > 1 - \delta$$

が成り立つ場合は，不要な属性として m を $cand(Y)$ から取り除くことができる．$cand(Y)$ からこの様な m を除外した後，残った属性に対してのみ Y からの探索枝を張ればよい．

4.2.2 形式概念の重複生成回避

先に触れた通り，同一の形式概念を抽出可能な生成元は一般に複数存在する．よって，抽出が不要であることが判明している形式概念の重複した生成は除去すべきである．ここでは，次の二つの観測をもとに，こうした重複生成を回避する方法について述べる．

観察 4.1
属性集合 $Y \subseteq M$ を考える．$\varphi(\psi(Y))\backslash Y$ 中の任意の属性 m について，$\varphi(\psi(Y \cup \{m\})) = \varphi(\psi(Y))$ である．つまり，$Y \cup \{m\}$ および Y のそれぞれからは，同一の形式概念が抽出される． ■

観測 4.1 より，生成元 Y と，$\varphi(\psi(Y))\backslash Y$ 中の任意の属性 m で拡張した Y の子供 $Y \cup \{m\}$ からは，同一の形式概念が抽出されることがわかる．つまり，$Y \cup \{m\}$ は δ-生成元には成り得ない．よって，Y の拡張候補集合 $cand(Y)$ から，こうした m はすべて除去可能である．

観察 4.2
$Y \subseteq M$ を属性集合，$Y \cup \{m\}$ を Y の子供とする．もし，$m' \prec m$ なる任意の属性 $m' \in \varphi(\psi(Y \cup \{m\}))\backslash\varphi(\psi(Y))$ が存在するならば，深さ優先探索において $Y \cup \{m\}$ より先

⋆1 同頻度の属性間に仮定可能な順序のバリエーションを考慮していないことから，厳密には，要請を満たすすべての生成元を調べていることにはなっていない．

Vol.2009-MPS-76 No.14
Vol.2009-BIO-19 No.14
2009/12/17

に処理される属性集合 Y' で，$\varphi(\psi(Y')) = \varphi(\psi(Y \cup \{b\}))$ となるものが必ず存在する[⋆1]．■

観測 4.2 より，$m' \prec m$ かつ $m' \in \varphi(\psi(Y \cup \{m\}))\backslash\varphi(\psi(Y))$ を満たす属性 m' が存在するならば，$Y \cup \{m\}$ から抽出される形式概念は，深さ優先探索の過程において，これまでのいずれかの生成元から抽出可能であったことがわかる．よって，暫定 Top-N リストに含まれる形式概念，あるいは，暫定リストから洩れた形式概念の中に，$\psi(Y \cup \{m\}))$ を包含する外延を有するものが存在する場合は，極大性の要請から，$Y \cup \{m\}$ の任意の拡張を枝刈り，直ちにバックトラックすることができる．なお，こうした形式概念が存在しない場合は，外延 $\psi(Y \cup \{m\}))$ を与えるそれ以前の生成元が，δ-生成元ではなかったことを意味している．

4.2.3 暫定 Top-N レア形式概念に基づく分枝限定

ここでのターゲットとなるレア形式概念は，その内包の δ-生成元の評価値 (一般性) が上位 N 以内のものである．よって，Top-N になれる見込みのない生成元はすべて探索の対象から除外することができる．この様な生成元は，探索中に見つかった暫定的な Top-N レア概念に基づいて同定することができる．

いま，生成元 Y の拡張候補集合 $cand(Y)$ から，上述した不要な候補を除いた候補集合をあらためて $cand(Y) = \{m_1, \ldots, m_k\}$ としよう．ここで，ある属性 $m_i \in cand(Y)$ により Y を拡張して得られる生成元 $Y' = Y \cup \{m_i\}$ を調べるとする．先に議論した通り，M 中の属性は，その頻度降順に順序が付けられ，各生成元中の属性もその順序に従う．よって，Y' の一般性は，$generality(Y') = sup(m)$ となる．いま，暫定 N 位のレア概念の生成元評価値を α とする．$sup(m_i) < \alpha$ である時，生成元 Y' が最終的に上位 N 位以内に入ることはあり得ない．また，$cand(Y)$ 中の m_i 以降の任意の属性 $m_j \in \{m_{i+1}, \ldots, m_k\}$ を用いて Y を拡張しても，生成元 $Y'' = Y \cup \{m_j\}$ の評価値は $sup(m_i)$ 以下，すなわち α 未満となり，やはり，最終的に上位 N 位以内に入る見込みがないことがわかる．さらに，一般性は，生成元の拡張に伴い単調に減少することから，Y' の任意の後者，および，Y'' の任意の後者についても，上位 N 位以内に入る見込みがないと結論できる．よって，$sup(m_i) < \alpha$ であることが観測された時点で，直ちに $cand(Y)$ 中の残りの属性 $m_{i+1}, \ldots, m_k$ による Y の拡張処理をすべて枝刈ることが可能となる．

4.3 抽出アルゴリズム

上述した枝刈り機構を組み込んだ Top-N レア形式概念抽出アルゴリズムの擬似コードを図 1 に示す．

⋆1 こうした m' は，文献 12) における**左候補** (*left candidate*) に相当するものである．

5. 実　　験

本章では，前章で提案したアルゴリズムに基づく実験システムにより得られた結果について述べる．なお，システムは C 言語で実装し，Intel Core2 Duo E9300 (1.2GHz)・主記憶 1GB の PC 環境で実行した．

5.1 データセット

実験には，Web 文書クラスタリングのベンチマークデータである *BankSearch* を用いた[7]．これは，“*CommercialBanks*”，“*BuildingSocieties*”，“*InsuranceAgencies*”，“*Java*”，“$C/C++$”，“*VisualBasic*”，“*Astronomy*”，“*Biology*”，“*Soccer*”，“*MotorSport*” および “*Sport*” の 11 カテゴリに属する Web 文書 (HTML テキスト) から成り，各カテゴリ毎に 1,000 文書を集めた，合計 11,000 の文書から成るデータセットである．

前処理として，まず，HTML タグを除去してプレーンテキスト化し，そこから，形容詞・副詞・ストップワードを除去した．ステミング処理[11] を行なった後，中間頻度の 3232 語彙を属性として抽出した．なお，カテゴリ情報は，各文書中に陽には現れない．

5.2 簡潔なレア形式概念の抽出例

ここでは，落差閾値 $\delta = 0.8$, レアネス閾値 $R = 10$ の設定のもとで Top-5 のレア概念抽出を行なった際に得られた概念の一例を示す．

外延:

{ A0045, A0596, B0401, B0539, B0628, B0642, H0054, H0295, J0001, J0464 }

δ-生成元：

{ update, condition, insurance, add }

含意属性：

{ rate, offer, note, interest }

外延は文書 ID の集合で表されており，先頭のアルファベットはカテゴリの別を表す．この概念は $A =$ “*CommercialBanks*”・$B =$ “*BuildingSocieties*”・$H =$ “*Biology*”・$J =$ “*MotorSport*” の四つのカテゴリにまたがり，かつ，頻度がおよそ 0.09% のレアなものである．すなわち，クロスオーバ概念の一例となっている．

生成元の一般性は 2,025，すなわち，生成元は，少なくとも全体のおよそ 18% の文書

Vol.2009-MPS-76 No.14
Vol.2009-BIO-19 No.14
2009/12/17

Input :
(G, M, I) : a formal context where
δ : a minimum reduction threshold
R : a rareness threshold
N : an integer for Top-N
Output :
$\mathcal{CRC}$: the set of Top-N maximal concise rare concepts

---

procedure main() :
$\mathcal{CRC} \leftarrow \phi$;
$current\_min = 0$;
Arrange the attibutes of M in support descending order ;
TopNCRCFind(ϕ, M, $\mathcal{CRC}$, 0) ;
return $\mathcal{CRC}$;

---

procedure TopNCRCFind(Y, $Cand$, $\mathcal{CRC}$) :
$Branch \leftarrow Cand \backslash \{m \in Cand \mid \frac{support(Y \cup \{m\})}{support(Y)} > 1 - \delta\}$;
for each $m \in Branch$ such that $tail(Y) \prec m$ in predefined order **do**
begin
if $\mathcal{CRC}$ tentatively contains N-th ones and
$generality(\varphi(\psi(Y))) < current\_min$ **then**
break;
endif
$CRC \leftarrow (Y \cup \{m\}, \psi(Y \cup \{m\}), \varphi(\psi(Y \cup \{m\})))$;
if $support((Y \cup \{m\})') \leq R$ **then**
if $\nexists CRC \in \mathcal{CRC}$ such that
whose extent properly subsumes $\psi(Y \cup \{m\})$ **then**
TopNListUpdate($\mathcal{CRC}$, CRC) ;
endif
else
TopNCRCFind($Y \cup \{m\}$, $Cand \backslash \{m\}$, $\mathcal{CRC}$) ;
endif
end

---

procedure TopNListUpdate($\mathcal{CRC}$, CRC) :
$\mathcal{CRC} \leftarrow \mathcal{CRC} \cup \{CRC\}$;
if $\mathcal{CRC}$ tentatively contains N-th ones **then**
$current\_min \leftarrow generality$-value of N-th generator ;
Remove M-th ones from $\mathcal{CRC}$ such that $N < M$;
endif

図 1 簡潔なレア形式概念の Top-N 抽出アルゴリズム

に出現する語彙から成っている．特に，insurance および condition が含まれていることから,『保険』の話題を含む文書群であろうとの推測ができるが，確かにどの文書にも保険 (insurance) に関わる話題が含まれていた．

含意属性とは，形式概念の内包から生成元を取り除いた残りの属性であり，生成元が含意する，すなわち，生成元に自然に付随する副次的な属性であり，外延を特徴付ける上で冗長なものと考えられる．こうした生成元と含意属性の明確な区別により，概念 (外延) の意味を解釈する際に考慮すべき属性 (語彙) が絞り込まれる．

なお，このレア概念を抽出するのに要した計算時間は，約 76 秒であった．

別の抽出例として，落差閾値 $\delta = 0.8$, レアネス閾値 $R = 10$ の設定のもとで Top-20 の抽出をした際に得られたレア概念のひとつを示す．

外延：

{ G0554, H0054, H0065, H0295, H0311, X0043, X0058 }

δ-生成元：

{ right, insure, add, resource }

含意属性：

{ list, system, show, offer, question, create, book, support, life, state, address,
fact, word, study, family, goal, force, help, promote, eye }

外延は，$G =$ “$Astronomy''$・$H =$ “$Biology''$・$X =$ “$Sport''$ の三つのカテゴリの文書群から成る頻度がおよそ 0.06% のレアなものである．ここでは特に，生成元と含意属性に含まれる語彙数の違いに注目したい．

生成元を構成するわずか 4 つの語彙が，20 の語彙を含意している．通常の形式概念を考える場合は，これら 24 の語彙をもとにして，この概念の意味解釈を試みるが，それは容易なことではない．しかし，ここでは生成元を構成する 4 つの語彙のみに注目できるので，その解釈がより明確になることが期待できるであろう．

なお，この時の計算時間は約 134 秒であった．

6. おわりに

レア概念 (パターン) の重要性は認識されているが，そうした概念の意味解釈は一般に容易ではない．こうした問題に対処すべく，本稿では，少数の一般的な属性で特徴付けられる

Vol.2009-MPS-76 No.14
Vol.2009-BIO-19 No.14
2009/12/17

概念の解釈は容易であるとの考えのもとに，簡潔なレア概念を定式化し，その Top-N 抽出アルゴリズムを与えた．計算機実験により，本手法により簡潔なクロスオーバ概念が抽出できることが確認できた．

一方，現在のアルゴリズムでは，探索対象となる生成元は，属性をその頻度降順に従って組み上げたもののみである．その意味で，可能な生成元には強い制約が課せられており，抽出可能なレア概念に限定が掛かかっている．さらに興味深いレア概念を抽出するためには，この制約の緩和が不可欠であり，今後考察を進めたいと考えている．

参 考 文 献

1) B.Ganter and R.Wille. Formal concept analysis - Mathematical foundations. Springer, 284 pages, 1999.
2) T. Uno, M. Kiyomi and H. Arimura. LCM ver. 2: Efficient mining algorithm for frequent/closed/maximal itemsets. Proc. of IEEE ICDM'04 Workshop - FIMI'04, http://sunsite.informatik.rwth-aachen.de/verb+Publications/CEUR-WS//Vol-126/, 2004.
3) G.Dong and J.Li, Efficient mining of emerging patterns: Discovering trends and differences, 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 43–52, 1999.
4) H. Alhammady and K. Ramamohanarao. Using emerging patterns and decision trees in rare-class classification, 4th IEEE International Conference on Data Mining, 315–318, IEEE Computer Society, 2004.
5) S.D. Bay, M.J.Pazzani. Detecting group differences: Mining contrast sets, Data Mining and Knowledge Discovery, 5, 213 – 246, Kluwer, 2001.
6) P.K. Novak and N. Lavrac, Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining The Journal of Machine Learning Research archive, Volume 10, 377 – 403, 2009.
7) M. P. Sinka and D. W. Corne, A large benchmark dataset for web document clustering, Soft Computing Systems: Design, Manegement and Applications, Series of Frontiers in Artificial Intelligence and Applications, 87, 881 – 890, 2002.
8) N. Pasquier, Y. Bastide, R. Taouil and L. Lakhal, Efficient mining of association rules using closed itemset lattices, Information Systems, 24(1), pp. 25 - 46, 1999.
9) E. Tomita and T. Kameda, An efficient branch-and-bound algorithm for finding a maximum clique with computational experiments", Journal of Global Optimization, 37, pp. 95 - 111, 2007.
10) E. Tomita and T. Seki, An efficient branch and bound algorithm for finding a maximum clique", Proceedings of the 4th International Conference on Discrete Mathematics and Theoretical Computer Science - DMTCS'03, LNCS-2731, 278 – 289 (2003).
11) M. F. Porter, An algorithm for suffix stripping", Program, 14(3), pp. 130 - 137, 1980.
12) A. Li, M. Haraguchi and Y. Okubo, Implicit groups of web pages as constrained top N concepts, Proc. of the 2008 IEEE/WIC/ACM Int'l Conf. on Web Intelligence and Intelligent Agent Technology Workshops, pp. 190 - 194, 2008.
13) M. Haraguchi and Y. Okubo, An extended branch-and-bound search algorithm for finding top-N formal concepts of documents", New Frontiers in Artificial Intelligence, JSAI 2006 Conference and Workshops, Tokyo, Japan, June 5-9, 2006, Revised Selected Papers, LNCS-4384, pp. 276 - 288, 2007.
14) M. Haraguchi and Y. Okubo, A method for pinpoint clustering of web pages with pseudo-clique search", Federation over the Web, International Workshop, Dagstuhl Castle, Germany, May 1 - 6, 2005, Revised Selected Papers, LNAI-3847, pp. 59 - 78, 2006.
15) Y. Okubo and M. Haraguchi, Finding conceptual document clusters with improved top-N formal concept search", Proc. of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence - WI'06, pp. 347 - 351, 2006.